lunedì 9 novembre 2020

FORMULARIO STATISTICA

Funzione

Definire una funzione ***y<-function(x)(3*x+7)***

Grafico

plot(y) se voglio dire gli estremi ***plot(y, -10, 10)***

Cancellare lo script

rm(list=ls())

Costruzione variabile nominale

Costruzione vettore nominale sesso di bambini (esempio)

sesso<-("M", "F", "M", "M", "M", "F")

Frequenza relativa

La frequenza di un valore divisa per il numero totale di frequenze

str(x) —> per vedere la stringa "x" ovvero tutti gli elementi

---

Tabella delle frequenze, frequenze relative o proporzionali

#tabella delle frequenze

freq_peso<-table(peso)

#tabella delle frequenze relative o proporzionali chiamata "rel peso"

rel_peso<- prop.table(freq_peso)

Approssimare alla seconda decimale

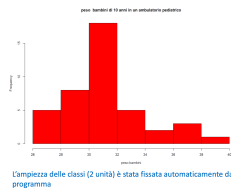
appr_peso<- round(rel_peso, digits=2)

Frequenze percentuali

#frequenze percentuali

freq_percent<- round(rel_peso*100, digits=2)

Istogrammi

hist(peso, col="green", main="freq assolute peso", xlab= "peso bambini", ylab= "frequenze assolute")

Altri grafici

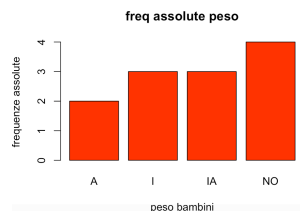
A) Diagrammi a torta

tabella<-table(allergie) tabella sulla console

pie(tabella)

B) Diagramma a barre

#diagramma a barre

barplot(tabella)

Calcolo della media

Media campionaria = $(x_1 + x_2 + \ldots x_n)/n$

peso_medio <- mean(peso)

Media ponderata = $(x_1 * f_1 + x_2 * f_2 + \ldots x_n * f_n)/n$

wa <- sum(x*f)/n

Mediana

median(peso)

Summery

Il comando "***summery()***" fornisce:

- gli estremi del campione;
- la mediana e i quartili;
- la media campionaria.

Varianza

Serve per determinare la dispersione dei dati rispetto alla media campionaria

$$s^2 = (x_k - media) * f/(n - 1)$$

var()

Deviazione standard

a) s2<-var() —> ***sqrt (s2)*** —> $\sqrt{Var}(x)$

b) ***sd ()***

Coefficiente di variazione

Serve per confrontare la deviazione standard e il valore assoluto della media campionaria

coeff_di_variazione <- sd() / mean()

Data frame

Se voglio leggere o vedere solo una colonna della tabella (ad esempio)

#leggere una variabile all'interno di un data set

inquinante$Lago.3 —> str(inquinante$Lago.3)

Percentile

è una misura usata in statistica per indicare il minimo valore sotto al quale ricade una data percentuale degli altri elementi sotto osservazione.

Es calcola il 10 percentuale

perc10<-quantile (x, 0.10) (0.10 perchè è 10%)

Tra i percentuali assumono particolare importanza i QUARTILI sono quei valori/ modalità che ripartiscono la popolazione in quattro parti di uguale.

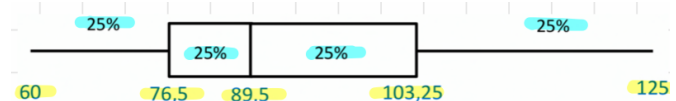
Ordinare il campione

sort()

Boxplot

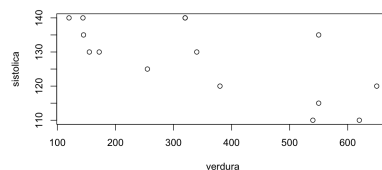
È un metodo grafico per rappresentare la suddivisione in quartili dei dati.

Si ottiene tracciando una linea che va dal più piccolo al più grande dei dati e due suddivisioni (box) che rappresentano i quartili.

Ogni box contiene il 25% dei dati:

boxplot (x, horizontal = TRUE, col="red", main="titolo")

Scatterplot

è un tipo di grafico in cui due variabili di un set di dati sono riportate su uno spazio cartesiano.

plot (x,y)

- se c'è qualche correlazione —> curva;
- se c'è correlazione lineare —> retta.

Covarianza

Per verificare se fra due variabili statistiche c'è qualche legame lineare.

$$cov(x, y) = (x_i - xmedia)(y_i - ymedia)/(n-1)$$

cov (x,y)

La covarianza può essere:

POSITIVA	NEGATIVA	NULLA
X e Y variano nella STESSA DIREZIONE	X e Y variano in DIREZIONI OPPOSTE	nessuna tendenza
al crescere di X cresce Y al diminuire di X diminuisce Y e viceversa	al crescere di X Y tende a diminuire	cov(X,Y)=0 ↘ NESSUNA CORRELAZIONE

lunedì 9 novembre 2020

Coefficiente di correlazione di Pearson

Serve per calcolare il grado di intensità della correlazione lineare di una coppia i campioni

cor (x,y) —> “r”

Valutazione sull’intensità di correlazione

A) valori positivi di r

0 < r < 0.25	debole o nessuna
0.25 < r < 0.75	moderata
0.75 < r < 1	forte

B) valori negativi di r

-0.25 < r < 0	debole o nessuna
-0.75 < r < -0.25	moderata
-1 < r < -0.75	forte

Nel caso in cui vi sia una significativa correlazione lineare (FORTE) si può costruire una retta chiamata retta di regressione

Retta di regressione

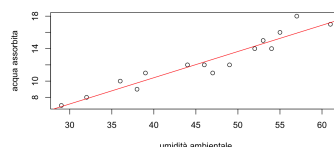
m = pendenza → β1 = $r_{xy} \frac{s_y}{s_x}$; q = intercetta → β0 = $\bar{y} - \hat{\beta}_1 \bar{x}$

Con R : ***lm(x~y)***

Sovrapposizione tra retta di regressione e i dati reali del campione

plot(x~y)

abline(retta di regressione)

lunedì 9 novembre 2020

Probabilità:

Spazio campionario

Simbolo: **Ω**

Significato: insieme di tutti i possibili esiti di un esperimento

---

Evento

Simbolo: **E**

Significato: sottoinsieme dello spazio campionario E ⊂ Ω

Operazioni tra insiemi

A) UNIONE **A ∪ B = { x ∈ Ω; x ∈ A oppure x ∈ B }**

B) INTERSEZIONE **A ∩ B = { x ∈ Ω; x ∈ A e x ∈ B}**

C) COMPLEMENTARE A^c **= { x ∈ Ω; x ∉ A}**

Leggi di de Morgan: $(A \cup B)^c = A^c \cap B^c$ e $(A \cap B)^c = A^c \cup B^c$

Insieme delle parti

Diciamo *"insieme delle parti"* una famiglia A di sottoinsiemi di Ω che soddisfi i seguenti criteri:

1. A non è vuoto —> $A \neq \varnothing$
2. Se E ∈ A, allora —> $E^c \in A$
3. Se $E_1, E_2 \in A$ allora —> $E_1 \cup E_2 \in A$ (A è chiuso rispetto alle operazioni di unione, intersezione)

P(Ω) = famiglia di tutti i sottoinsiemi di Ω / insieme delle parti di Ω

Definizione di probabilità

Dati Ω spazio campionario e A sistema di eventi, diciamo *probabilità su Ω* un'applicazione —> **P : A → R; E → P (E)**

che soddisfa: 1. 0 < P (E) < 1

2. P (Ω) = 1

3. Se A, B ∈ A, A ∩ B = Ø allora $P(A \cup B) = P(A) + P(B)$

Regole del calcolo della probabilità

A) Probabilità del complementare **P(A^c) = 1 - P(A)**

B) Evento impossibile **P(Ø) = 0**

C) Partizione dell'evento certo $P(B) = P(B \cap A) + P(B \cap A^c)$

D) Ordinamento $A \subset C \rightarrow P(A) \leq P(B)$

E) Unione di eventi non disgiunti $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Eventi indipendenti

A e B sono indipendenti se il verificarsi di A non influenza la probabilità che si verifichi B e viceversa

$$P(A \cap B) = P(A)P(B)$$

Spazi di probabilità uniformi

Diciamo che (Ω,P) è uno *"spazio di probabilità uniforme"* se

1. Ω è finito;
2. ∀ ω ∈ Ω, P { ω }) = p indipendentemente dal valore di ω. (Ovvero se la probabilità è la stessa per ogni esito in Ω)

Usiamo la notazione -| A | = tutti gli elementi dell'insieme A

- p = 1 / | Ω |

Dunque ∀ A ∈ P(Ω) abbiamo che P(A) = (|A|)*p = | A | / | Ω |

Dunque in questo caso si può dire che Probabilità = casi favorevoli/casi possibili

---

Variabili aleatorie

Data una popolazione generica Ω, diciamo *variabile aleatoria su Ω* una variabile "X" che assume casualmente i suoi valori nella popolazione Ω.

Generalmente possiamo considerare l'evento *X = k*:

Considerata una variabile aleatoria X, ha senso calcolare *P (X=k)*

Variabile aleatoria discreta

È una variabile aleatoria che assume valori discreti x=0,1,2,…n e y= $x_1, x_2, x_3 \ldots$

Sempre in generale diciamo che x_k ∈ range di X se P (x=x_k) ≠ 0

Densità di probabilità o funzione di distribuzione o pdf (probability distribuition function)

Funzione : fornisce la probabilità di ogni evento costruito a partire da X :

$$f_x(x) = P(X=x) \text{ oppure } p_k = P(X=k)$$

Regole :

1. ∀ k $0<p_k<1$
2. p_k = P(X=k) = 0 se k ∉ range di X
3. $\sum p_k = 1$

Valore atteso

In generale il valore atteso di una variabile casuale discreta (che assuma cioè solo un numero finito o una infinità numerabile di valori) è dato dalla somma dei possibili valori di tale variabile, ciascuno moltiplicato per la probabilità di essere assunto (ossia di verificarsi), cioè è la media ponderata dei possibili risultati.

Diciamo “valore atteso” di X (media, valor medio, speranza matematica, attesa) la quantità:

$$\mu = E[X] = \sum x_i * P(X=x_i)$$

Proprietà:

1. E[cX] = cE[X]
2. E[X + Y] = E[X] + E[Y]

Varianza e deviazione standard di una v.a.

Rappresentano la dispersione dei possibili valori di X rispetto al valore atteso.

Data X, variabile aleatoria finita, diciamo “varianza di X” la quantità:

$$\sigma\text{^}2 = s^2 = Var(X) = \sum (x_i - \mu)\text{^}2 * P(X=x_i)$$

Diciamo “deviazione standard di X”:

$$\sigma = s = \sqrt{Var(X)}$$

Proprietà:

1. Var (aX) = a^2Var(X) ;
2. Var (a + X) = Var (X) (invarianza per traslazione);
3. Se X,Y sono indipendenti allora E[XY]=E[X] E[Y] e Var(X+Y)=Var(X) Var(Y).

lunedì 9 novembre 2020

Processi di Bernoulli - Distribuzione Binomiale

Prova Bernoulliana

Diciamo prova bernoulliana un esperimento che può avere solo due risultati

V successo	P(V) = p
F insuccesso	P(F) = 1 - p

N.B: p per convenzione indica sempre la probabilità di successo

Variabile aleatoria di Bernoulli

Diciamo v.a. di Bernoulli la v.a. :

$$X \sim B(p)$$

Essa può assumere due valori:

X = 1	Successo (V)	P (X=1) = p(1) = p
X = 0	Insuccesso (F)	P (X=0) = p(0) 0 1-p

Valore atteso

Indica quante volte in media abbiamo successo

$$E[X] = p$$

Varianza

$$Var(X) = p(1 - p)$$

Processi di Bernoulli

È una sequenza (anche illimitata) di prove bernoulliane.

1. Prove indipendenti;
2. Tutte con la stessa probabilità di successo p ∊ (0,1).

- **Limitati** : sequenza di numero **fissato** n di prove bernoulliane ;
- **Illimitati** : sequenza di numero **infinito** di provare bernoulliane .

lunedì 9 novembre 2020

Processi di Bernoulli limitati :

- *X=* : numero di successi ottenuti in n prove;
- X può assumere i valori: 0,1,2,..n ;
- X è detta variabile aleatoria binomiale di parametri n (numero di prove) e p (probabilità di successo in ogni prova)
- Si indica $X \sim B(n, p)$

Valore atteso e varianza di un processo bernoulliano limitato

Valore atteso : $E[X] = np$; Varianza : $Var(X) = np(1-p)$

FUNZIONI IN R :

Fattoriale

FUNZIONE: ***n!= factorial(n)***

Coefficiente binomiale

Calcoliamo coefficienti binomiali

FUNZIONE:

A) Regola matematica: ***n binomiale k —> choose(n,k): k<=n***

B) Funzione precostituita: ***PP_7<- dbinom (7, size=n, prob=p)***

Per calcolare ad esempio P(x<=2) = P(x=0)+P(x=1)+P(x=2)

usiamo la funzione ***p<-sum(dbinom((0:2), size=n, prob=p))***

Setnames

Questa è una funzione comoda che imposta i nomi su un oggetto e restituisce l'oggetto. È molto utile alla fine della definizione di una funzione in cui si sta creando l'oggetto da restituire e si preferisce non memorizzarlo con un nome solo per poter assegnare i nomi.

In questo caso

Tabuliamo pdf:

tab.pdf<-setNames(pdf,k)

pdf = distribuzione e k = vettore range della variabile

Realizzazione di una variabile

Assumiamo di aver estratto concretamente un valore x dalla popolazione:

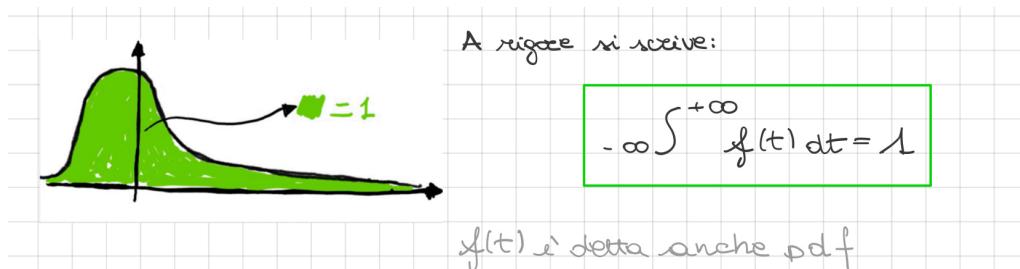
x è detto realizzazione della variabile aleatoria X

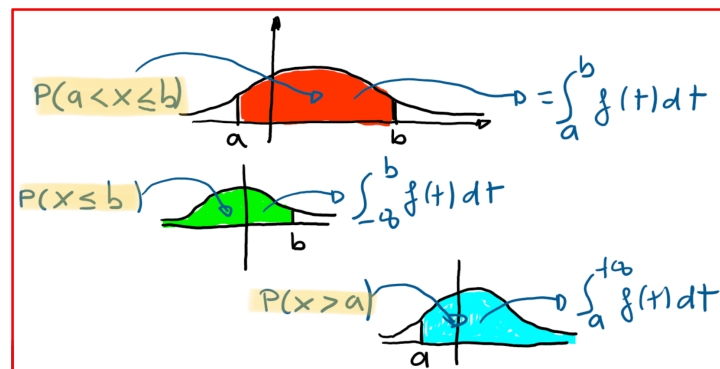
Es: scommettiamo sul lancio di due dati: scommettiamo sul 7 ma all'estrazione esce il numero 5 (realizzazione di X). Dato di fatto: abbiamo perso qualsiasi fossero le previsioni probabilistiche di X

Variabile aleatoria discreta

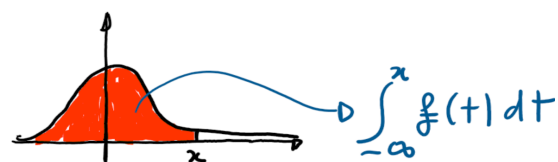
Assumono valori in un intervallo di numeri reali (es. [0,∞], [0.4,10] ecc)

Funzione densità PDF

È una funzione 0<f(t)<1 tale che l'area sottesa al suo grafico sia pari a 1:

Calcolo della probabilità

$\forall a, b \in R, a < b$ la probabilità $P(a < X < b)$ è data dall'area sottesa al grafico tra l'intervallo $[a, b]$:

Funzione di ripartizione CDF (cumulative distribuition function)

$$F(t) = P(X < x)$$

1. *F funzione non decrescente;*
2. $\lim_{x \to -\infty} F(x) = 0;$
3. $\lim_{x \to +\infty} F(x) = 1.$

Funzione di ripartizione e calcolo della probabilità

A) Dato un intervallo [a,b] $P(a < X < b) = F(b) - F(a)$

B) La funzione di ripartizione ci permette di calcolare $P(a < X < b)$, dunque $F(x)$ descrive completamente la distribuzione della v.a. X.

C) Calcolare gli estremi o non calcolarli non fa differenza

Valore attesa e varianza per una variabile continua

Valore atteso: μ = E[X] = $\int_{-\infty}^{+\infty} t\, f(t)dt$

Intuitivamente può essere considerato come la media dei possibile valori di X pesati in modo continuo dalla densità f(t).

Varianza: σ^2= Var(X) = E[(x-μ)^2] = $\int_{-\infty}^{+\infty} (t-\mu)^2 f(t)dt$

Distribuzione normale standard

Z ~ N (0,1)

E[Z] = 0 e Var(Z)=1

$$\varphi(t) := \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$$

In R : ***φ(x) = dnorm(x, mean=0, sd=1)***

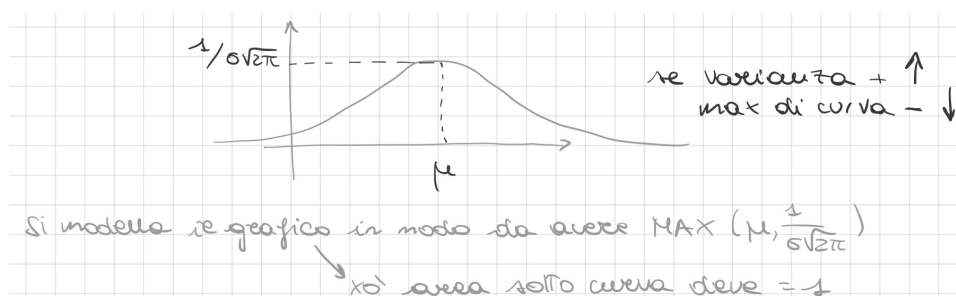
Calcolo della probabilità per distribuzione Normale Standard

Utilizziamo R perchè non si può fare in altro modo.

Distribuzione normale di media μ e varianza σ^2

X ~ N (μ, σ^2) definita da X = σZ + μ

X ~ N(μ, σ^2) $\iff Z \sim N(0,1)$

Curva della distribuzione normale:

Proprietà della distr. normale

Date:

- X v.a. normale di media μx e varianza σx^2, **X ~ N(μx,σx^2)**
- Y v.a. normale di media μy e varianza σy^2, **Y ~ N(μy,σy^2)**
- X e Y indipendenti

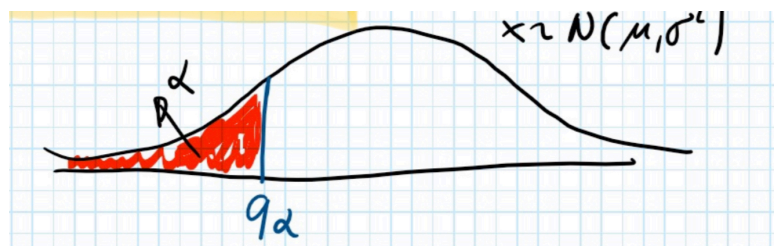
ALLORA: **X + Y ~ N (μx+μy , σx^2+σy^2)**

Attenzione! Le varianze di sommano, le deviazioni standard no!

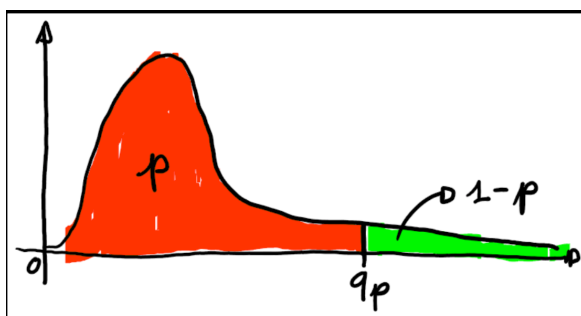
Quantili della distr. normale standard

Considerata la v.a.normale standard $Z \sim N(0,1)$ e fissata una probabilità $0 < \alpha < 1$

Diciamo *quantile di ordine* α, il valore z_a tale che: **P (Z < z_a) = α**

In R : ***qnorm(alpha, mean=0, sd=1) = qnorm(α)***

Quantili in generale

Data una v.a. X di densità f(t) e funzione F (t) = P(X<t), fissata una probabilità 0<p<1 diciamo *quantile di ordine p:*

il valore $q_p \in R$ tale che $P(X < q_p) = p$

Le proprietà dei quantili z_p della distribuzione normale standard sono basate sulla simmetria rispetto all'asse delle ordinate della funzione gaussiana.

Dunque non sono valide in generale per i quantili q_p una distribuzione qualsiasi

Tabella CODICI R per distribuzioni

Distribuzione	**pdf (densità)**	**cdf (ripartizione)**	**quantili di ordine α**
X ~ unif [a,b]	***dunif*** (x, min=a, max=b)	***punif*** ((x, min=a, max=b)	***qunif*** ((x, min=a, max=b)
X ~ N (μ, σ^2)	***dnorm*** (x, mean=μ, sd=σ)	***pnorm*** (x, mean=μ, sd=σ)	***qnorm*** (x, mean=μ, sd=σ)
X ~ B (n,p)	***dbinom*** (x, size=n, prob=p)	***pbinom*** (x, size=n, prob=p)	***qbinom*** (x, size=n, prob=p)

Per riassumere:

pdf (densità)	***d “nome” = (x, parametri)***
cdf (funzione di ripartizione)	***p “nome” = (x, parametri)***
Quantili di ordine α	***Q “nome” = (x, parametri)***

lunedì 9 novembre 2020

Statistica inferenziale

Vogliamo ottenere informazioni sull'intera popolazione a partire dai dati ottenuti attraverso il campionamento.

Procedimento

1) Fissiamo un campione aleatorio
2) Dallo studio teorico otteniamo informazioni sulla distribuzione di probabilità
3) Realizzazione:

- estraiamo campione di dati;
- Con strumenti teorici (B) effettuiamo stime su μ e σ

Campione aleatorio **indipendente identicamente distribuito**

Chiamiamo *campione aleatorio* ***i.i.d.*** una famiglia di variabili aleatorie che soddisfino:

1. Indipendenti
2. Stessa distribuzione (= pdf)
3. Stessa varianza (= σ^2)
4. Stesso valore atteso (= E[X])

Valore atteso E[Xmedia]

E [Xmedia] = μ

Non dipende da n (ampiezza)

Varianza Var(Xmedia)

σ^2 / n

Dipende da n (↑n ; ↓Var)

Se la popolazione è normale di media μ e varianza σ^2

Xmedia ~ N (μ , σ^2 / n)

lunedì 9 novembre 2020

Teorema del limite centrale

Se:

- campione i.i.d.
- media μ
- varianza σ^2
- **n >= 30**

Allora: Assumiamo che la popolazione abbia **distribuzione normale:**

Xmedia ~~ N (μ , σ^2 / n)

Realizzazione di un campione aleatorio

Un campione casuale semplice estratto concretamente dalla popolazione

Su R comando **r+nome variabile(…)**

N (μ , σ^2)	rnorm (n, mean=mu, sd=sigma)
unif ([a,b])	runif (n, max=b, min=a)
B (n,p)	rbinom (n, size=N, prob=p)

Intervallo di confidenza per media nota con varianza non nota

a) Estraiamo un campione casuale di ampiezza n

b) Fissiamo un livello di fiducia/confidenza CL=1 - α

cl<-0.99

alpha<-1-cl

alpha

c) Calcoliamo la media campionaria

xbar=mean(x)

d) Intervallo di confidenza per la media, al livello di fiducia 1-α è **[xbar-E, xbar+E]**

zstar<-qnorm(1-alpha/2)

E<-zstar * sigma/sqrt(n)

IC<- xbar+c(-1,+1)*E

Risposta: - l'intervallo di c. è [estremo sx, estremo dx]

- il livello medio di rumorosità è μ = xbar ± E

Se la popolazione ha **media e varianza entrambe NON note?**

σ^2 non nota -> non utilizzabile per costruire intervallo di confidenza

Introduciamo quindi:

- ***Varianza campionaria*** $S^2 = 1/(n-1)^* \Sigma(x_j - xbar)^2$
- ***Deviazione standard campionaria*** S= $\sqrt{S^2}$

Esse sono chiamate *"variabili aleatorie statistiche"* -> servono per stimare μ e σ.

Si verifica che ***xbar-μ / (S / √n) = t_n-1 =*** distribuzione t di student con n-1 gradi di libertà

Distribuzione t-di student X~t_n

Descritta da densità f(t) con grafico molto simile alla normale standard (campana simmetrico rispetto ad origine) ma è più precisa della normale

E[X]=0

Var(X)= n / n-2 -> sempre Var(X)>1

In R: ***f(t)= dt(t, df=n)***

Quantili della t di student

t* = qt(alpha, df=n)

Intervallo di confidenza per media nota con varianza non nota

a) consideriamo un campione casuale di ampiezza n
b) Calcoliamo media e varianza campionarie xbar<-mean(x) s^2<- var(x)
c) Intervallo di confidenza al livello di confidenza CL= 1-alpha

[xbar-E , xbar+E]

E<-t*S/sqrt(n)

t* = qt (alpha, df=n)

Intervallo di confidenza è dunque: μ **= xbar ± E**

$\mu \in$ **[estremo sx; estremo dx]**

Proporzione di successi in una popolazione bernoulliana

- Consideriamo la popolazione bernoulliana (successo/insuccesso) con probabilità di successo p in ogni prova (p ignota)
- Consideriamo un campione aleatorio i.i.d.
- Variabile aleatoria che conta il numero di successi del campione:

ṗ = (numero di successi di campione di ampiezza n) / n

SE n*p>=5 e n*(1-p)>=5 → ALLORA **[(ṗ-p) / sqrt(p*(1-p)/n)] ~ N**

(Ha distribuzione normale standard)

Intervalli di confidenza per proporzione di successi in pop. Bernoulliana

- popolazione Bernoulliana
- Fissiamo livello di confidenza CL = 1-alpha
- Estraiamo campione casuale
- Calcoliamo ṗ
- Verifichiamo che n*p>=5 e n*(1-p)>=5
- Calcoliamo errore statistico **E= z_(1-alpha/2) * [ṗ(1-p) / n]**

z_(1-alpha/2) = qnorm(1-alpha/2)

Intervallo di confidenza è dunque: **[ṗ-E ; ṗ+E]**

p = ṗ±E

Come si stima la varianza di una popolazione?

Stima della varianza

- Consideriamo una popolazione con varianza NON nota

SE consideriamo la varianza campionaria S^2 → ALLORA si verifica che **E[S^2]=σ^2**

S^2 è un buon stimatore di σ^2

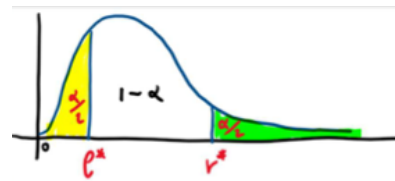
Si verifica inoltre che la variabile aleatoria ***(n-1)* * S^2/σ^2** ha distribuzione **"chi quadro con n-1 gradi di libertà"**

Distribuzione chi-quadro con n-1 gradi di libertà (χ)

In R: ***dchiq (x, df=n)***

Quantili della chi-quadro

Sinistro : ***lstar<- qchisq(alpha/2, df=n-1)***

Destro : ***rstar<-qchisq(1-alpha/2, df=n-1)***

Intervalli di confidenza per la varianza

- Popolazione normale con varianza NON nota
- Fissiamo il livello di confidenza CL = 1-alpha
- Estraiamo campione casuale
- Media e varianza campionarie xbar e S^2
- Calcoliamo i quantili lstar e rstar

Intervallo di confidenza è dunque: $\left(\frac{(n-1)s^2}{r^*} ; \frac{(n-1)s^2}{l^*}\right)$

Test di ipotesi

Procedura per assumere con un certo livello di significatività statistica se l'affermazione H_0 debba essere rifiutata o non rifiutata.

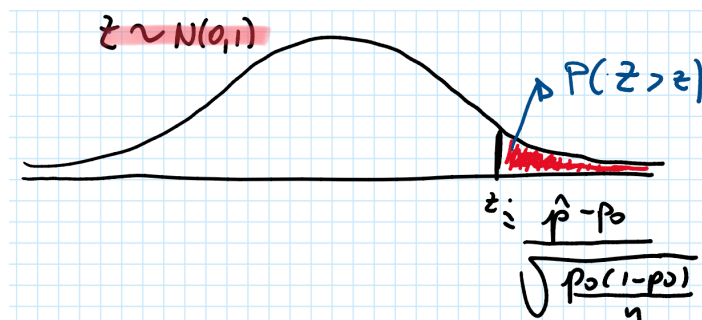
A) Ipotesi nulla H_0

Afferma che un parametro della popolazione è uguale ad un valore teorico fissato.

B) Ipotesi alternativa H_A

È l'intervallo di valori che il parametro assume quando l'ipotesi nulla è FALSA.

Come stabilire se rifiutare o no H_0?

Statistica di test z

P(Z>z) = p-value (Z=normale standard; z=statistica di test)

$\alpha \in$ (0,1) = livello di significatività

(In generale $0.01<\alpha<0.1$ → tra l'1% e il 10%)

Quindi:

SE	ALLORA
p-value<=α	Rifiutiamo H_0
p-value>α	Non rifiutiamo H_0

Test di ipotesi sulla proporzione p di successi in una popolazione Bernoulliana

- Popolazione bernoulliana con p ∊ (0,1)
- Fissiamo un *valore di confronto* p_0 e un *livello di significatività* α
- Estraiamo un campione in modo che ***n****p_0 **> 5 e *n**(1-**p_0**) > 5**
- Calcoliamo la proporzione campionaria ***ṗ = num di successi / n***
- Calcoliamo statistica di test ***z* = (*ṗ* -** p_0**) / *sqrt[(*** p_0 **(1-**p_0**) / *n]***

Test ad una coda a destra	Test a due code	Test ad una coda a sinistra
H_0: p=p_0 ; H_A: p>p_0	H_0: p=p_0 ; H_A: p≠p_0	H_0: p=p_0 ; H_A: p<p_0
p-value = P(Z>z)	p-value = 2P(Z>|z|)	p-value = P(Z<z)

prop.test(x,n,p=p0,alternative="greater");

prop.test(x,n,p=p0,alternative="two.sided");

prop.test(x,n,p=p0,alternative="less").

Statistica di test per la costruzione di processi decisionali

Tipologia di test	Statistica di test	Distribuzioni delle statistica di test
Test sulla popolazione H_0: p=p_0	***z* = (*ṗ* -** p_0**) / *sqrt[(***p_0 **(1-**p_0**) / *n]***	N (0,1)
Test sulla media H_0: mu=mu0	a) sigma nota ***z= (xbar-mu0) / (sigma/ sqrt(n))*** b) sigma non nota ***T= (xbar-mu0) / (s/sqrt(n))***	a) sigma nota N (0,1) b) sigma non nota t_(n-1)
Test sulla deviazione standard H_0: sigma=sigma0	***chiquadro= [(n-1)**** S^2***] /*** $sigma0^2$	chiquadro^2 con n-1 gradi di libertà

lunedì 9 novembre 2020

Test di ipotesi per media e varianza

Test per la media con varianza NOTA

- *Popolazione con media μ incognita e σ^2 nota*
- *Fissiamo H_0: mu=mu0*
- *Fissiamo α (0,1)*
- *Estraiamo campione con ampiezza n*
- Calcoliamo media campionaria xbar = mean(x)
- Costruiamo z = (xbar-mu0) / (sigma/sqrt(n))

Test ad una coda a destra	Test a due code	Test ad una coda a sinistra
H_0: mu=mu0 ; H_A: mu>mu0	H_0: mu=mu0 ; H_A: mu≠mu0	H_0: mu=mu0 ; H_A: mu<mu0
p-value = P(Z>z)	p-value = 2P(Z>|z|)	p-value = P(Z<z)

p-value<-pnorm(z) ; p-value<-2*pnorm(z) ; p-value<-1-pnorm(z) .

prop.test(x,n,p=p0,alternative="greater");

prop.test(x,n,p=p0,alternative="two.sided");

prop.test(x,n,p=p0,alternative="less").

Test per la media con varianza NON NOTA

(primi 4 punti uguali a prima)

- Calcoliamo media campionaria xbar e varianza campionaria S^2
- Costruiamo t = (xbar-mu0) / (s/sqrt(n))

Test ad una coda a destra	Test a due code	Test ad una coda a sinistra
H_0: mu=mu0 ; H_A: mu>mu0	H_0: mu=mu0 ; H_A: mu≠mu0	H_0: mu=mu0 ; H_A: mu<mu0
p-value = P(T>t)	p-value = 2P(T>|t|)	p-value = P(T<t)

p-value<-pt(t, df=n-1) ; p-value<-2*pt(abs(t), df=n-1, lower.tail=FALSE ;

p-value<-1-pt(t, df=n-1) .

oppure

t.test(x,mu=mu0, alternative="greater");
t.test(x,mu=mu0,alternative="two.sided");
t.test(x,mu=mu0,alternative="less")

Test per la varianza

- Popolazione normale con σ^2 incognita
- Fissiamo H_0: sigma^2=sigma0^2 e fissiamo α (0,1)
- Estraiamo campione con ampiezza n
- Calcoliamo varianza campionaria S^2
- Costruiamo chiquadro = [(n-1)*S^2] / sigma0^2

Test ad una coda a destra	Test a due code	Test ad una coda a sinistra
H_A: sigma>sigma0	H_A: sigma≠sigma0	H_A: sigma<sigma0
p-value = P(X>χ^2)	p-value = 2min[P(X<χ^2); P(X>χ^2)]	p-value = P(X<χ^2)

p-value<-pchisq(chi, df=n-1, lower.tail=FALSE) ;

p-value<-2*min[pchisq(chi,df=n-1,lower.tail=TRUE), pchisq(chi,df=n-1,lower.tail=FALSE)

p-value<-pchisq(chi,df=n-1,lower.tail=TRUE)

[Test non parametrici: test di wilcoxon per la mediana

- Consideriamo H_0: mediana=m
- Fissiamo α (0,1)
- Estraiamo campione casuale

Test ad una coda a destra	Test a due code	Test ad una coda a sinistra
H_A: mediana>m	H_A: mediana≠m	H_A: mediana<m
wilcox.test(x, mu=m, alternative="greater")	***wilcox.test(x, mu=m, alternative="two.sided")***	***wilcox.test(x,mu=m, alternative="less")***

Il test non accetta TIES, ovvero ripetizioni nel vettore di dati.

Si aggiunge "exact=FALSE", ovvero R non calcola il p-value preciso, ma lo stima in modo che possa funzionare lo stesso. **]**

lunedì 9 novembre 2020

Confronto tra parametri (media, dev.standard etc) tra due popolazioni diverse

Test sul confronto tra due popolazioni bernoulliane

- Abbiamo popolazione 1 con *proporzione* p_1 di successo e

  popolazione 2 con *proporzione* p_2 di successo

- Ampiezza dei campioni n_1 e n_2, non per forza uguali
- Estraiamo campioni X e Y
- Calcoliamo ***phat1=ṗ1=numero successi in X/***n_1 e

  phat2=ṗ2=numero successi in Y/n_2

- Costruiamo la statistica di test così:
  - **ṗ=**proporzione totale di successi=***(phat1****n_1 **+ *phat2****n_2 ***) /*** n_1+n_2
  - Statistica di test=***z= (ṗ1-ṗ2) / sqrt[ṗ*(1-ṗ)*(1/***n_1***+1/***n_2***)]***

    Si distribuisce come una normale standard N(0,1)

Codifica R:

- Vettore numero di successi ***vp= c (n1*phat1, n2*phat2)***
- Vettore ampiezze dei campioni ***vn= (n1, n2)***
- Effettuiamo il prop.test ***prop.test(vp, vn, alternative="greater, two.sided, less")***

lunedì 9 novembre 2020

Test di confronto tra medie, mediane e varianze

Test di confronto tra medie di due popolazioni

- Abbiamo popolazione X e popolazione Y
- Ampiezza dei campioni n_x e n_y, non per forza uguali
- Estraiamo campioni X e Y
- Consideriamo medie e varianze REALI mu_x, mu_y, $sigma2_x$, $sigma2_y$
- H_0: mu_x=mu_y
- Calcoliamo xbar, ybar (medie campionarie)
- Calcoliamo S_x^2 e S_y^2 varianze campionarie

Distinguiamo i test sulla base delle informazioni su $sigma2_x$ e $sigma2_y$

1) Varianze note

Statistica di test: z = (xbar-ybar) / sqrt[$sigma2_x/n_x + sigma2_y/n_y$]

Coda a dx: ***p-value<-pnorm(z,lower.tail=FALSE);***

Due code: ***p-value<-2*pnorm(abs(z), lower.tail=FALSE);***

Coda a sx: ***p-value<-1-pnorm(z, lower.tail=TRUE).***

2) Varianze NON note ma uguali $sigma2_x = sigma2_y$

Statistica di test: t = (xbar-ybar) / sqrt[$1/n_x+1/n_y$]

Coda a dx: ***t.test<-(x, y, alternative="greater", var.equal=TRUE)***

Due code: ***t.test(x, y, alternative="two.sided", var.equal=TRUE)***

Coda a sx: ***t.test(x, y, alternative="less",var.equal=TRUE)***

3) Varianze NON note e ignote $sigma2_x \neq sigma2_y$

Statistica di test: t = (xbar-ybar) / sqrt[$S_x^2/n_x + S_y^2/n_y$]

Coda a dx: ***t.test<-(x, y, alternative="greater", var.equal=FALSE)***

Due code: ***t.test(x, y, alternative="two.sided", var.equal=FALSE)***

Coda a sx: ***t.test(x, y, alternative="less",var.equal=FALSE)***

Test di confronto tra mediane di due popolazioni

- H_0: medianaA = medianaB

Utilizziamo il test di Wilcoxon:

Coda a dx: ***wilcox.test<-(x, y, alternative="greater")***

Due code: ***wilcox.test(x, y, alternative="two.sided")***

Coda a sx: ***wilcox.test(x, y, alternative="less")***

Test di confronto tra varianze

- H_0: $sigma2_A = sigma2_B$
- Calcoliamo varianze campionarie S_A^2 e S_B^2

Statistica di test: f = S_A^2/S_B^2 distribuzione di Fisher

[Distribuzione di Fisher

Codifica R:

- Quantili: qf (α, n, m)
- P(F<f) = pf(α, n, m, lower.tail=TRUE)
- P(F>f) = pf(α, n, m, lower.tail=FALSE) **]**

Coda a dx: ***var.test<-(x, y, alternative="greater")***

Due code: ***var.test(x, y, alternative="two.sided")***

Coda a sx: ***var.test(x, y, alternative="less")***

lunedì 9 novembre 2020

Test di indipendenza

Serve per verificare se due variabili sono indipendenti o dipendono l'una dall'altra.

- H_0: variabili sono indipendenti
- H_A: variabili non indipendenti

Codifica R:

- Calcolare la matrice x=(25,13; 12,15) ***x<-rbind(c(25,13), c(12,15))***
- Eseguiamo il test chiquadro ***chisq.test(x)***

Attenzione: se le frequenze in una casella sono meno di 5 appare un Warning

allora ***chisq.test(x,simulate.p.value=TRUE)***

Test di adattamento ad un modello

Serve per verificare se un modello probabilistico ipotizzato per la popolazione sia o meno compatibile con i dati campionari che abbiamo.

Test chi-quadro di adattamento

- n categorie A1, A2 etc
- Vettore delle proporzioni teoriche di elementi per categoria è

  pt = p1,p2, …pk (p1+p2+…pk=1)
- Estraiamo campione casuale
- Frequenze per categoria z=z1,z2,…zk

Eseguiamo il test: ***chisq.test(z,p=pt)***

Test chi-quadro per le distribuzioni continue non note

Consideriamo popolazione con distribuzione X non nota

Serve per verificare se la popolazione si adatta ad un modello di distribuzione X_0

H_0: X=X_0 e H_A: X≠X_0

Si usa il **Test di Kolmogorov-Smirnov**: ***ks.test(x, "pchisq"/"pnorm"/"punif", df=n)***

Confronto tra le distribuzioni di due popolazioni

- Abbiamo popolazione X e popolazione Y

Serve per stabilire se le distribuzione delle due popolazioni sono uguali oppure diverse.

H_0: X=Y e H_A: X≠Y

- Estraiamo i campioni x e y

Eseguiamo il test: ***ks.test(x,y)***

Verifica della normalità di una popolazione

- Abbiamo popolazione X

Serve per stabilire se la distribuzione di probabilità della popolazione X è normale

H_0: X ha distribuzione normale; H_A: X non ha distribuzione normale

- Estraiamo un campione casuale x

Eseguiamo il **Test di Shapiro-Wilk**: ***shapiro.test(x)***

lunedì 9 novembre 2020

Test ANOVA (Analysis of variance)

È un test di ipotesi sulle medie che utilizza la varianza delle distribuzioni per arrivare al calcolo del p-value.

- Consideriamo n popolazioni indipendenti: $X_1, X_2, \dots X_n$
- Assumiamo che in tutte le popolazioni:
  - Distribuzione normale
  - Varianze tutte uguali (tutte la stessa varianza)

H_0: mu1=mu2=…muk ; H_A: almeno una ≠ dalle altre

- Estraiamo i campioni: x_1, x_2, x_n

- Costruiamo una lista con tutti i campioni: ***l<-list(maggio<-may, …)***
- Costruiamo il dataframe: ***d<-stack(l)***
- Guardiamo struttura del dataframe: ***str(d)***

Eseguiamo il test: ***oneway.test (values~ind, data=d, var.equal=TRUE)***

Se NON possiamo assumere la normalità della distribuzione?

Test di Kruskall-Wallis

- Consideriamo n popolazioni indipendenti, tutte con la stessa distribuzione (non necessariamente nota

H_0: mu1=mu2=…muk ; H_A: almeno una ≠ dalle altre

- Eseguiamo le altre operazioni descritte precedentemente

Eseguiamo il test: ***kruskal.test (values~ind, data=d)***